Black-box Stein Divergence Minimization For Learning Latent Variable Models

Chao Ma
Department of Computer Science
University College London
chao.ma.16@ucl.ac.uk

David Barber
Department of Computer Science
University College London
d.barber@cs.ucl.ac.uk

Abstract

Typical variational methods for learning latent variable models require tractable likelihood evaluation of both joint model $p(\mathbf{x}, \mathbf{z})$ and an approximate inference model $q(\mathbf{z}|\mathbf{x})$, which becomes the main constraint for many applications. To relax these requirements, by leveraging inclusive Stein divergence and generalizing it to handle latent variables, we propose Blackbox Stein Divergence Minimization (BBS), a new variational framework for learning doubly intractable latent variable models, coupled with implicit inference models with intractable likelihoods. Fully black box training now becomes possible, using only a little analytic information. As an example, truncated MCMC as implicit inference models are trained end-to-end while learning rectified Gaussian latent variable models.

1 Introduction

Confronting the needs for learning the underlying structures from data in an unsupervised way, we are encountering more and more complicated probabilistic models where learning and inference are difficult. Generally, a generative model with latent variables are defined as follows:

$$p(\mathbf{x}, \mathbf{z}|\theta) = p(\mathbf{x}|\mathbf{z}, \theta)p(\mathbf{z}) \tag{1}$$

Where $\mathbf{x}$ is the observation data and $\mathbf{z}$ is latent variable. Intractabilities may arise due to intractable normalization functions $p(\mathbf{x})$ (for posterior density $p(\mathbf{z}|\mathbf{x}, \theta)$) and $\mathcal{Z}(\theta)$ (for joint density $p(\mathbf{x}, \mathbf{z}|\theta)$), i.e., the *doubly intractable distributions* [18] that cover a wide range of useful probabilistic models.

A number of algorithms for such distributions have been studied [16, 18]. However, they mainly focus on different problem settings, i.e. inferring $\mathbf{z}$ with intractable normalizer $\mathcal{Z}(\mathbf{z})$ of forward model $p(\mathbf{x}|\mathbf{z})$, a special case of intractable $\mathcal{Z}(\theta)$ for joint density $p(\mathbf{x}, \mathbf{z}|\theta)$. When it comes to parameter learning, the EM algorithm is not feasible due to the intractable normalizer $\mathcal{Z}(\theta)$ of the joint density. A notable work [22] generalized the score matching approach for learning doubly-intractable latent variable models. However, it is only constrained to exponential family distributions and relies on the use of MCMC samples, which are unbiased only in asymptotes and are challenging to scale.

Instead of an MCMC E-step, a Variational-Autoencoder-like structure may be considered, where an auxiliary inference model $q(\mathbf{z}|\mathbf{x}, \lambda)$ is used to perform approximate inference on $\mathbf{z}$. The following Variational Free Energy (VFE) are optimized wrt. variational parameters λ and model parameters θ:

$$\log p(\mathbf{x}|\theta) = -\mathrm{KL}(p_{data}||p_\theta) + \text{const.} \geq \mathcal{F}_{VFE} = \langle \log p(\mathbf{x}, \mathbf{z}|\theta) - \log q(\mathbf{z}|\mathbf{x}, \lambda) \rangle_q \tag{2}$$

MC based black-box variational methods has been introduced to to give estimation of gradients of VFE wrt. λ [7, 17, 19, 20]. However, these methods are still not fully black box due to the need for complete access to the likelihood of inference model, which becomes the main constraint for proposing expressive inference models. This motivates us to expand our scope and also tackle another intractability referred as *wild variational approximation* [9, 12], i.e. fully black box inference models $q(\mathbf{z}|\mathbf{x}, \lambda)$ that are intractable to evaluate, but easy to get unbiased samples from.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

In this paper, we will provide a learning and inference framework for general latent variable models in the face of doubly intractable joint densities and wild variational inference models. We propose to optimize a consistent estimator of a new distance measure called inclusive Stein divergence, in place of KL-divergence used in Eq.(2). We expect our approach to enable us to design more flexible and powerful models without worrying about intractabilities.

2 Inclusive Stein Divergence for learning latent variable models

In previous works [12, 15, 23], exclusive Stein divergence based objectives for variational inference have drawn strong interests recently, due to their ability to evaluate the distance between model and samples, and to handle unnormalized distributions[1]. However, the use of exclusive Stein divergence suffers from the following drawbacks: firstly, it needs model parameter θ to be fixed, which is not suitable for ML learning; secondly, it does not handle latent variables.

In contrast to exclusive Stein divergence based methods, here we leverage inclusive Stein divergence to generalize stein variational framework to be able to perform ML learning with latent variables, under doubly intractable joint distribution and wild inference model. Assume we have a set of unbiased samples $\mathcal{D} = \{\mathbf{x}_i\}$ drawn from unknown ground truth distribution $p^\star(\mathbf{x})$. We learn our generative model $p_\theta(\mathbf{x}) = \sum_{\mathbf{z}} p_\theta(\mathbf{x}, \mathbf{z})$ by minimizing the inclusive Stein divergence $\mathbb{S}(p^\star||p_\theta)$:

$$\theta^\star = \operatorname*{argmin}_\theta \mathbb{S}(p^\star||p_\theta)$$

If the theoretical minimum of $\mathbb{S}(p^\star||p_\theta) = 0$ is achieved, one will have $p^\star = p_{\theta^\star}$. To perform gradient descent learning, we calculate the gradient $\nabla_\theta \frac{1}{2}\mathbb{S}(p^\star||p_\theta)$:

$$\nabla_\theta \frac{1}{2}\mathbb{S}(p^\star||p_\theta) = \mathbb{D}(p^\star||p_\theta)\nabla_\theta \max_\phi \sum_i \mathcal{T}_{p_\theta}\phi(\mathbf{x}_i) \tag{3}$$

$$= \mathbb{D}(p^\star||p_\theta)\nabla_\theta \max_\phi \sum_i \frac{1}{p(\mathbf{x}_i|\theta)} \left\langle \frac{p(\mathbf{x}_i, \mathbf{z}|\theta)}{q(\mathbf{z}|\mathbf{x}, \lambda, \theta)} \partial_\epsilon \log p_{\theta,[T^{-1}_{\epsilon,\phi}]}(\mathbf{x}_i, \mathbf{z})|_{\epsilon=0} \right\rangle_{q_{\lambda,\theta}} \tag{4}$$

Where $\mathcal{T}_p \cdot = \langle \nabla_{\mathbf{x}} \log p, \cdot \rangle + \langle \nabla_{\mathbf{x}}, \cdot \rangle$ is the Stein operator. T is defined as an infinitesimal distortion on $\mathbf{x}$: $T_{\epsilon,\phi}(\mathbf{x}) = \mathbf{x} + \epsilon\phi(\mathbf{x})$, and $p_{\theta,[T^{-1}_{\epsilon,\phi}]}(\mathbf{y})$ is the distribution of $\mathbf{y} = T^{-1}_{\epsilon,\phi}(\mathbf{x}), \mathbf{x} \sim p_\theta(\mathbf{x})$. Likewise, $p_{\theta,[T^{-1}_{\epsilon,\phi}]}(\mathbf{y}_1, \mathbf{y}_2)$ is the joint distribution of $(\mathbf{y}_1, \mathbf{y}_2) = T^{-1}_{\epsilon,\phi}(\mathbf{x}, \mathbf{z}), (\mathbf{x}, \mathbf{z}) \sim p_\theta(\mathbf{x}, \mathbf{z})$, and $T_{\epsilon,\phi}(\mathbf{x}, \mathbf{z}) = (\mathbf{x} + \epsilon\phi(\mathbf{x}), \mathbf{z})$. Derivation of (4) is available in Appendix B. Now an auxiliary approximate inference model $q(\mathbf{z}|\mathbf{x}, \lambda, \theta)$ is incorporated to perform importance sampling, with weights defined by $w_{ij}(\mathbf{z}_j) = \frac{p(\mathbf{x}_i, \mathbf{z}_j|\theta)}{q(\mathbf{z}_j|\mathbf{x}, \lambda)}$. Therefore, we can see that the corporation of auxiliary approximate inference model $q(\mathbf{z}|\mathbf{x}, \lambda, \theta)$ allows us to handle latent variables under Stein divergence.

3 Black-box gradients computable under doubly intractable p and wild q

3.1 Black-box Stein Importance Sampling for $w_{ij}(\mathbf{z}_j)$ evaluation

The major obstacle for applying inclusive Stein divergence for learning latent variable models is the estimation of ratio $w_{ij}(\mathbf{z}_j) = \frac{p(\mathbf{x}_i, \mathbf{z}_j|\theta)}{q(\mathbf{z}_j|\mathbf{x}, \lambda, \theta)}$. In recent work of GANs [4] and DHIMs [21], the main strategy is to train a separate neural network as a ratio estimator to estimate *unnormalized* $w_{ij}(\mathbf{z}_j)$ by classifying whether the sample comes from $p(\mathbf{x}_i, \mathbf{z}_j|\theta)$ or $q(\mathbf{z}_j|\mathbf{x}, \lambda, \theta)$. Here, we avoid the need for a separate neural network, by using an alternative method based on the importance sampling like structure in inclusive Stein divergence framework (4). First note that using MC samples $\mathbf{z}_j \sim q(\mathbf{z}|\mathbf{x}, \lambda)$ we can arrive at an approximated version of the objective:

$$\max_\phi \sum_i \frac{1}{p(\mathbf{x}_i|\theta)} \left\langle \frac{p(\mathbf{x}_i, \mathbf{z}|\theta)}{q(\mathbf{z}|\mathbf{x}, \lambda, \theta)} \partial_\epsilon \log p_{\theta,[T^{-1}_{\epsilon,\phi}]}(\mathbf{x}_i, \mathbf{z})|_{\epsilon=0} \right\rangle_{q(\mathbf{z}|\mathbf{x}, \lambda, \theta)}$$

$$\simeq \max_\phi \sum_i \frac{1}{\sum_j w_{ij}} \sum_j w_{ij} \partial_\epsilon \log p_{\theta,[T^{-1}_{\epsilon,\phi}]}(\mathbf{x}_i, \mathbf{z}_j)|_{\epsilon=0}$$

In the above approximation of inclusive Stein divergence, the problem of estimating Stein divergence (and related derivatives) reduces to the estimation of *normalized* importance weights $\widetilde{w}_{ij} = \frac{w_{ij}}{\sum_j w_{ij}}$,

[1] We briefly review the basic definitions and results necessary for this paper in supplementary materials. For a complete introduction, readers might refer to [1, 11].

using only information of score function of $p(\mathbf{x}_i, \mathbf{z}_j|\theta)$ wrt. latent variables $\mathbf{z}$, and samples $\mathbf{z}_j$ from $q(\mathbf{z}_j|\mathbf{x}, \lambda, \theta)$. Fortunately, efficient techniques satisfying these requirements has been proposed recently also based on Stein divergence, i.e. Blackbox Stein importance sampling (BBIS) [13]. BBIS has demonstrated its capacity of being both accurate (in terms of nice convergence property) and efficient (in terms of simplicity of algorithm). The computational complexity of BBIS is $\mathcal{O}(MJ^2)$, where M is the size of minibatch and J is the number of importance samples per likelihood term. With J being small (typically 2 to 10), BBIS (as a quadratic optimization problem) can be solved fast using off-the-shelf optimizers [8]. Compared with exact IS, BBIS has the additional benefit of improving the estimation accuracy and reducing variance [13].

3.2 Learning signal $\mathbb{D}(p^\star||p_\theta)$ evaluation

The special structure of inclusive Stein divergence allows us to run black-box gradient estimations given only unnormalized joint $f(\mathbf{x}_i, \mathbf{z}_j|\theta)$ and unbiased samples from $q(\mathbf{z}|\mathbf{x}, \lambda)$. Before we proceed, first note that in the gradient (4) if we restrict ϕ to belong to the function class of RKHS in a unit ball, then we can express optimal $\phi^\star_{p_\theta}(\mathbf{x})$ analytically [1, 14]:

$$\phi^\star_{p_\theta}(\mathbf{x}) \propto \left\langle \mathcal{T}_{p(\mathbf{x}'|\theta)} k(\mathbf{x}', \mathbf{x}) \right\rangle_{\mathbf{x}' \sim p^\star}$$

Where $k(\mathbf{x}', \mathbf{x})$ is a specific kernel function that we choose. With the same set of importance sampling weights we can give estimate of score function $\nabla_{\mathbf{x}_i} \log p(\mathbf{x}_i|\theta)$ at data points $\mathbf{x}_i$:

$$\nabla_{\mathbf{x}_i} \log p(\mathbf{x}_i|\theta) \approx \frac{1}{\sum_j w_{ij}} \sum_j w_{ij} \nabla_{\mathbf{x}_i} \log p(\mathbf{x}, \mathbf{z}_j|\theta)$$

Which can be used to give an estimation of $\phi^\star_{p_\theta}(\mathbf{x})$.

3.3 Black box gradient evaluations

Given the special structure of inclusive Stein divergence, there are many possible options to evaluate gradients in a black-box way. Assuming that the inference model $q(\mathbf{z}_j|\mathbf{x}, \lambda, \theta)$ is reparameterizable, i.e. $\mathbf{z} = g_{\lambda,\theta}(\xi, \mathbf{x})$ for some deterministic, continuous and differentiable function g and auxiliary random variable $\xi \sim p(\xi)$. Then we can directly perform gradient descent on empirical estimation of inclusive Stein divergence

$$\nabla_\theta \max_\phi \sum_i \partial_\epsilon \log p_{\theta,[T^{-1}_{\epsilon,\phi}]}(\mathbf{x}_i)|_{\epsilon=0} \tag{5}$$

$$\simeq \nabla_\theta \sum_i \frac{1}{\sum_j w_{ij}} \sum_j w_{ij} \mathcal{T}_{p_\theta(\mathbf{x}|\mathbf{z})} \phi^\star_{p_\theta}(\mathbf{x}) \tag{6}$$

The fact that $\mathcal{T}_{p_\theta(\mathbf{x}|\mathbf{z})}\phi = \partial_\epsilon \log p_{\theta,[T^{-1}_{\epsilon,\phi}]}(\mathbf{x}_i, \mathbf{z})|_{\epsilon=0}$ can be derived by running again the same argument as shown in Appendix B. This can be then decomposed in to gradients $\nabla_\theta \mathcal{T}_{p_\theta(\mathbf{x}|\mathbf{z})} \phi^\star_{p_\theta}(\mathbf{x})$, $\nabla_\theta \mathbf{z}$ and $\partial_\theta w_{ij}$. $\nabla_\theta \mathbf{z}$ is easy to compute provided that $\mathbf{z} \sim q$ is reparameterizable.

Note that, $\mathcal{T}_{p_\theta(\mathbf{x}|\mathbf{z})}$ depends on θ only through unnormalized joint $f(\mathbf{x}_i, \mathbf{z}_j|\theta)$, therefore terms involving $\nabla_\theta \mathcal{T}_{p_\theta(\mathbf{x}|\mathbf{z})} \phi^\star_{p_\theta}(\mathbf{x})$ can be easily computed using automatic differentiation. Also, weights w_{ij} depends on θ only through $f_{ij} = f(\mathbf{x}_i, \mathbf{z}_j|\theta)$, therefore can be computed using chain rule

$$\partial_\theta w_{ij} = \left\langle \nabla_{f_{i,\cdot}} w_{ij}, \partial_\theta f_{i,\cdot} \right\rangle \tag{7}$$

Where $\partial_\theta f_{ij}$ can be computed analytically using automatic differentiation, and $\partial_{f_{ij}} w_{ij}$ can be easily estimated given only a minibatch of pairs of input-output measurements $\{f_{i,\cdot}, w_{ij}\}$ (e.g. [2, 3, 10, 24]). In practice, we found that replacing f_{ij} by its score functions $\nabla_{\mathbf{x}} \log f_{ij}$ and $\nabla_{\mathbf{z}} \log f_{ij}$ for black-box gradient estimation will help convergence of learning. Gradient evaluations for variational parameters λ can also be conducted in a similar manner. Finally, we are able to present each iteration of the proposed black box Stein divergence minimization algorithm as follows, in Algorithm 1.

4 Experiments

We present experiments on inference and learning tasks on Rectified Gaussian latent variable models (RGLVM) to demonstrate the feasibility of our proposed method. The RGLVM is defined as:

$$p_{RG}(\mathbf{z}) \propto \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{\Sigma}) \prod \Theta(z_l), \;\; p_{RG}(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|W\mathbf{z}, \sigma^2\mathbf{I})$$

Algorithm 1 Learning with Black-box Stein divergence minimization (one update iteration)

1: Sampling: a minibatch $\mathcal{B} \subset \mathcal{D}$, and $\mathbf{z}^1, ..., \mathbf{z}^J \sim q(\mathbf{z}|\mathbf{x}_i, \lambda, \theta)$ for all $i \in \mathcal{B}$
2: Compute $\widetilde{\mathbf{w}}^\star$ by BBIS [13]
3: Using automatic differentiation tools, perform SGD on Stein divergence in Eq.(4) wrt. θ, λ based on approximated objectives (e.g.,Eq (6)) respectively

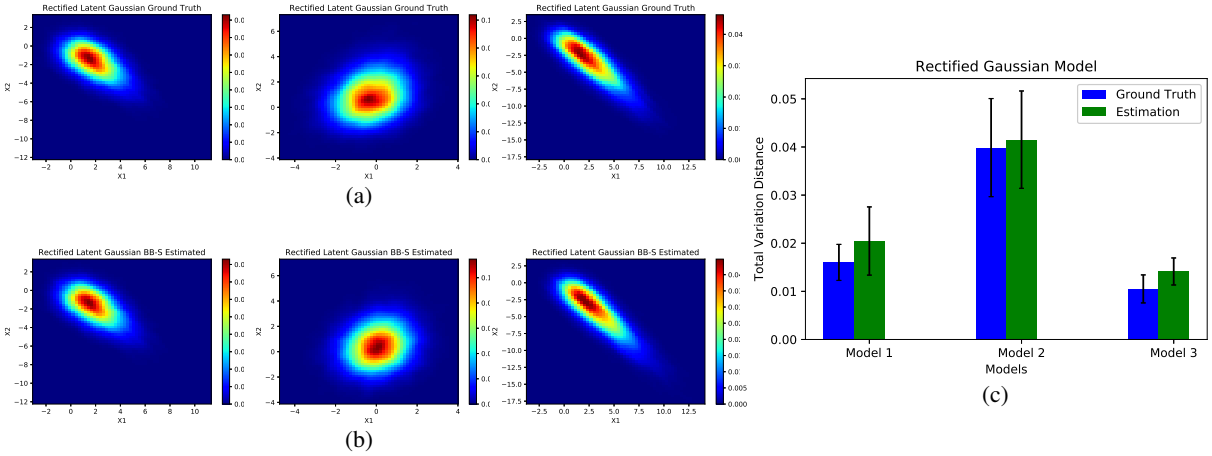

Figure 1: (a): Heat maps of ground truth RGLVMs; (b) Heat maps of learned RGLVMs by our method; (c): Total variation distances from the ground truth parameters (blue), and from the estimated model parameters (green)

In general, the normalizer for the joint RGLVM cannot be computed analytically, therefore the inference and learning problem is challenging for variational methods.

We use Langevin Dynamics as end-to-end truncated MCMC for our inference model:

$$\mathbf{z}_{t+1} = m_t(\mathbf{z}_t|\mathbf{x}, \epsilon, \theta, \eta_t) = \mathbf{z}_t + \frac{\epsilon_t}{2}\nabla_{\mathbf{z}} \log p(\mathbf{x}, \mathbf{z}|\theta) + \eta_t, \;\; \eta_t \sim \mathcal{N}(0, \epsilon\mathbf{I})$$

Where ϵ_t is the step size scalar at layer t, that serves as variational parameters. Such approximate MCMC network is flexible since information regarding the model is considered, but intractable to evaluate. We now run experiments to learn both the model parameters of 2-D RGLVMs depicted Fig1(a) and variational parameters$\{\epsilon_t\}$ simultaneously in an end-to-end manner, using black box inclusive Stein divergence minimization with $J = 2$. As shown in Fig 1(b), the proposed method can recover marginal distributions of original models with nearly indistinguishable approximations.

To evaluate the quality of estimation given by our method, we use total variation as the distance measure between data distributions based on empirical estimate of the following score of quality:

$$\delta(P, Q) = \sup_{\mathbf{x}} |P(\mathbf{x}) - Q(\mathbf{x})|, \;\; S(\hat{\theta}) = \delta(P_{data}, Q_{\hat{\theta}}), \;\; S(\theta^\star) = \delta(P_{data}, Q_{\theta^\star})$$

Where $\hat{\theta}$ and $\theta^\star$ are the estimated and ground truth parameter, respectively. Then we can compare whether the quality of estimation $S(\hat{\theta})$ is close to the quality of ground truth estimation $S(\theta^\star)$, as depicted in Fig 1(c). Indeed, the proposed blackbox inclusive Stein divergence minimization method is able to converge to fairly good solutions of the RGLVM problem.

5 Conclusion and Future Work

In this extended abstract, we have generalized inclusive Stein divergence to handle general doubly-intractable latent variable models. Under this framework, we are able to perform learning and inference, without knowing the normalization terms of the generative model (i.e. doubly intractable latent variable models), and the exact distribution of inference model. We've also proposed to train truncated MCMC samplers as implicit inference models in a purely end-to-end manner, that allows joint learning of model parameters and variational parameters. In the future, we will continue to investigate novel approaches of gradient estimation, conduct extensive numerical experiments on a number of datasets, and explore more real-life applications of our approach.

Acknowledgments Chao Ma thanks the China Scholarship Council and NSF of Guangdong Province of China (2015A030313574 and 2017A030313397) for funding his research.

References

[1] Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2016.

[2] Kris De Brabanter, Jos De Brabanter, Bart De Moor, and Irène Gijbels. Derivative estimation with local polynomial fitting. *Journal of Machine Learning Research*, 14(1):281–301, 2013.

[3] Jianqing Fan and Irene Gijbels. *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, volume 66. CRC Press, 1996.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[5] Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(Apr):695–709, 2005.

[6] Oliver Thomas Johnson. *Information theory and the central limit theorem*. World Scientific, 2004.

[7] Diederik P Kingma and Max Welling. Auto-encoding Variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[8] Dieter Kraft. A software package for sequential quadratic programming. *Forschungsbericht-Deutsche Forschungs- und Versuchsanstalt fur Luft- und Raumfahrt*, 1988.

[9] Yingzhen Li and Qiang Liu. Wild variational approximations. *NIPS Workshop in Advances in Approximate Bayesian Inference*, 2016.

[10] Yingzhen Li and Richard E Turner. Gradient estimators for implicit models. *NIPS Workshop in Advances in Approximate Bayesian Inference*, 2016.

[11] Qiang Liu. A short note on kernelized Stein discrepancy.

[12] Qiang Liu and Yihao Feng. Two methods for wild variational inference. *arXiv preprint arXiv:1612.00081*, 2016.

[13] Qiang Liu and Jason D Lee. Black-box importance sampling. *arXiv preprint arXiv:1610.05247*, 2016.

[14] Qiang Liu, Jason D Lee, and Michael I Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2016.

[15] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances In Neural Information Processing Systems*, pages 2370–2378, 2016.

[16] Anne-Marie Lyne, Mark Girolami, Yves Atchade, Heiko Strathmann, Daniel Simpson, et al. On russian roulette estimates for Bayesian inference with doubly-intractable likelihoods. *Statistical science*, 30(4):443–467, 2015.

[17] Andriy Mnih and Karol Gregor. Neural variational inference and learning in belief networks. *arXiv preprint arXiv:1402.0030*, 2014.

[18] Iain Murray, Zoubin Ghahramani, and David MacKay. MCMC for doubly-intractable distributions. *arXiv preprint arXiv:1206.6848*, 2012.

[19] Rajesh Ranganath, Sean Gerrish, and David M Blei. Black box variational inference. In *AISTATS*, pages 814–822, 2014.

[20] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.

[21] Dustin Tran, Rajesh Ranganath, and David M Blei. Deep and hierarchical implicit models. *arXiv preprint arXiv:1702.08896*, 2017.

[22] Eszter Vértes, UCL Gatsby Unit, and Maneesh Sahani. Learning doubly intractable latent variable models via score matching.

[23] Dilin Wang and Qiang Liu. Learning to draw samples: With application to amortized mle for generative adversarial learning. *arXiv preprint arXiv:1611.01722*, 2016.

[24] Shanggang Zhou and Douglas A Wolfe. On derivative estimation in spline regression. *Statistica Sinica*, pages 93–108, 2000.

Appendix

A Stein's Divergence

Stein divergence based probabilistic machine learning techniques have drawn strong interests from the community recently due to being able to naturally handle unnormalized distributions. Here we briefly review the basic definitions and results necessary for this paper, without proofs. For complete introduction readers might refer to [1, 11]. Stein's discrepancy $\mathbb{D}(q||p)$ between distributions $q(\mathbf{x})$ and $p(\mathbf{x})$ (without latent variables)is defined by:

$$\mathbb{D}(q||p) = \max_{\phi \in \mathcal{G}} \langle \mathcal{T}_p \phi(\mathbf{x}) \rangle_{\mathbf{x} \sim q}$$

Where $\phi(\mathbf{x})$ is a vector-valued test function with the same dimensionality as $\mathbf{x}$, that belongs to a restrictive set of function $\mathcal{G}$ that we choose. $\mathcal{T}_p$ is the Stein operator associated with distribution p defined as:

$$\mathcal{T}_p \phi = \langle \nabla_{\mathbf{x}} \log p, \phi \rangle + \langle \nabla_{\mathbf{x}}, \phi \rangle$$

Then we can define Stein divergence to be

$$\mathbb{S}(q||p) = \mathbb{D}^2(q||p) \geq 0$$

Therefore we can see that $\mathbb{S}(q||p)$ depends on p only through its score function $s_p(\mathbf{x}) = \nabla_{\mathbf{x}} \log p$, thus being able to handle unnormalized distribution p. By using stokes theorem, and assuming $p\phi$ vanishes on the boundary of $\mathbf{x} \in \mathcal{X}$, one can show that $\mathbb{S}(q||p) = 0$ if and only if $q = p$. Given this property, a number of works have propose to leverage Stein divergence for inference by minimizing Stein's divergence [12, 15, 23].

In practice the we may take the set of functions $\mathcal{G}$ in the optimization problem **??** to be the unit ball of vector-valued reproducing kernel Hilbert spaces (RKHS) with kernel $k(\mathbf{x}, \mathbf{x}')$:

$$\mathcal{G} = \{\phi \in \mathcal{H}^d | \|\phi\|_{\mathcal{H}^d} \leq 1\}$$

Then the solution under optimal ϕ can be written as:

$$\begin{aligned} \mathbb{D}(q||p) &= \max_{\phi \in \mathcal{G}} \langle \mathcal{T}_p \phi(\mathbf{x}) \rangle_{\mathbf{x} \sim q} \\ &= \sqrt{\langle k_p(\mathbf{x}, \mathbf{x}') \rangle_{\mathbf{x}, \mathbf{x}' \sim q}} \end{aligned}$$

Where k_p is the *Steinized* kernel of $k(\mathbf{x}, \mathbf{x}')$:

$$\begin{aligned} k_p(\mathbf{x}, \mathbf{x}') &= \mathcal{T}_p^{\mathbf{x}'}(\mathcal{T}_p^{\mathbf{x}} \otimes k(\mathbf{x}, \mathbf{x}')) \\ &= s_p(\mathbf{x})^\top k(\mathbf{x}, \mathbf{x}') s_p(\mathbf{x}') + s_p(\mathbf{x})^\top \nabla_{\mathbf{x}'} k(\mathbf{x}, \mathbf{x}') \\ &+ \nabla_{\mathbf{x}} k(\mathbf{x}, \mathbf{x}')^\top s_p(\mathbf{x}') + \operatorname{Tr} \nabla_{\mathbf{x}, \mathbf{x}} k(\mathbf{x}, \mathbf{x}') \end{aligned}$$

Therefore we can define the *Kernelized* Stein Divergence as:

$$\mathbb{S}(q||p) = \langle k_p(\mathbf{x}, \mathbf{x}') \rangle_{\mathbf{x}, \mathbf{x}' \sim q}$$

A more intuitive way of defining kernelized Stein divergence is given by [14]

$$\mathbb{S}(q||p) = \left\langle \delta_{p,q}(\mathbf{x})^\top k(\mathbf{x}, \mathbf{x}') \delta_{p,q}(\mathbf{x}) \right\rangle_{\mathbf{x}, \mathbf{x}' \sim q}$$

Where $\delta_{p,q}(\mathbf{x}) = s_p(\mathbf{x}) - s_q(\mathbf{x})$ is the difference of score functions between q and p. The know clear that kernelized Stein divergence is actually a kernelized version of score matching objective [5], i.e. the Fisher divergence [6]

$$\mathbb{F}(q||p) = \left\langle \|\delta_{p,q}(\mathbf{x})\|_2^2 \right\rangle_{\mathbf{x} \sim q}$$

B Derivation of Eq.(4)

We start by showing

$$\begin{aligned}\nabla_\theta \frac{1}{2}\mathbb{S}(p^\star||p_\theta) &\simeq \mathbb{D}(p^\star||p_\theta)\nabla_\theta \max_\phi \sum_i \mathcal{T}_{p_\theta}\phi(\mathbf{x}_i) \\ &= \mathbb{D}(p^\star||p_\theta)\nabla_\theta \max_\phi \sum_i \partial_\epsilon \log p_{\theta,[T^{-1}_{\epsilon,\phi}]}(\mathbf{x}_i)|_{\epsilon=0}\end{aligned}$$

The second equality of can be derived as follows:

$$\begin{aligned}\left(\partial_\epsilon \log p_{\theta,[T^{-1}_{\epsilon,\phi}]}(\mathbf{x}_i)\right)|_{\epsilon=0} &= (\partial_\epsilon \log p_\theta(T_{\epsilon,\phi}(\mathbf{x}_i))|\det \nabla_{\mathbf{x}_i} T_{\epsilon,\phi}(\mathbf{x}_i)|)\,|_{\epsilon=0} \\ &= \langle \nabla_{\mathbf{x}_i}\log p(T_{\epsilon,\phi}(\mathbf{x}_i)), \partial_\epsilon T_{\epsilon,\phi}(\mathbf{x}_i)\rangle\,|_{\epsilon=0} \\ &+ \operatorname{trace}\left(\nabla_{\mathbf{x}_i} T_{\epsilon,\phi}(\mathbf{x}_i)^{-1}\partial_\epsilon \nabla_{\mathbf{x}_i} T_{\epsilon,\phi}(\mathbf{x}_i)\right)|_{\epsilon=0} \\ &= \mathcal{T}_{p_\theta}\phi(\mathbf{x}_i)\end{aligned}$$

The first equality follows from transformation of random variable, the second term of second equality follows from matrix derivative of log determinant, and the third equality follows by noticing $T_{\epsilon,\phi}(\mathbf{x}) = \mathbf{x} + \epsilon\phi(\mathbf{x})$, $T_{\epsilon=0,\phi}(\mathbf{x}) = \mathbf{x}$, $\partial_\epsilon T_{\epsilon,\phi}(\mathbf{x}) = \phi(\mathbf{x})$, $\nabla_{\mathbf{x}} T_{\epsilon,\phi}(\mathbf{x}) = I$, and $\partial_\epsilon \nabla_{\mathbf{x}} T_{\epsilon,\phi}(\mathbf{x}) = \nabla_{\mathbf{x}}\phi(\mathbf{x})$.

Intuitively, this gives us an alternative perspective: minimizing $\mathbb{S}(p^\star||p_\theta)$ is equivalent to minimizing the magnitude of increase of model log likelihood under a infinitesimal perturbation vector $-\phi(\mathbf{x})$, along which is chosen to maximumly increase the log likelihood, multiplied by a learning signal $\mathbb{D}(p^\star||p_\theta)$. However from this perspective, though true, it is not easy to see the fact that when objective $\mathbb{S}(q||p)$ is minimized to zero, one will actually obtain $p_{\theta^\star} = p^\star$.

Now we proceed with derivation (ignoring $\mathbb{D}(p^\star||p_\theta)$ for now):

$$\begin{aligned}&\nabla_\theta \max_\phi \sum_i \partial_\epsilon \log p_{\theta,[T^{-1}_{\epsilon,\phi}]}(\mathbf{x}_i)|_{\epsilon=0} \\ &= \nabla_\theta \max_\phi \sum_i \frac{1}{p(\mathbf{x}|\theta)}\partial_\epsilon \sum_{\mathbf{z}} p_{\theta,[T^{-1}_{\epsilon,\phi}]}(\mathbf{x}_i,\mathbf{z})|_{\epsilon=0} \\ &= \nabla_\theta \max_\phi \sum_i \frac{1}{p(\mathbf{x}|\theta)} \sum_{\mathbf{z}} p(\mathbf{x},\mathbf{z}|\theta)\partial_\epsilon \log p_{\theta,[T^{-1}_{\epsilon,\phi}]}(\mathbf{x}_i,\mathbf{z})|_{\epsilon=0} \\ &= \nabla_\theta \max_\phi \sum_i \frac{1}{p(\mathbf{x}|\theta)} \sum_{\mathbf{z}} q(\mathbf{z}|\mathbf{x},\lambda,\theta)\frac{p(\mathbf{x},\mathbf{z}|\theta)}{q(\mathbf{z}|\mathbf{x},\lambda,\theta)}\partial_\epsilon \log p_{\theta,[T^{-1}_{\epsilon,\phi}]}(\mathbf{x}_i,\mathbf{z})|_{\epsilon=0} \\ &= \nabla_\theta \max_\phi \sum_i \frac{1}{p(\mathbf{x}_i|\theta)} \left\langle \frac{p(\mathbf{x}_i,\mathbf{z}|\theta)}{q(\mathbf{z}|\mathbf{x},\lambda,\theta)}\partial_\epsilon \log p_{\theta,[T^{-1}_{\epsilon,\phi}]}(\mathbf{x}_i,\mathbf{z})|_{\epsilon=0}\right\rangle_{q(\mathbf{z}|\mathbf{x},\lambda,\theta)}\end{aligned}$$

Which completes the derivation.